# Uma Abordagem de Componentes Combinados para a Geração de Funções de Ordenação usando Programação Genética

**Humberto Mossri de Almeida**[1], **Marcos Andre Gonçalves**[1]

[1] Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627, Belo Horizonte, MG, CEP 31270-010, Brasil

{hmossri,mgoncalv}@dcc.ufmg.br

***Abstract.*** *Due to the advent of the Web, the information retrieval task has become a very complex and challenging problem. Because of that, several ranking functions have been investigated throughout the years. However, most of them attempt to be very general in nature, i.e., they were designed to be effective in any type of collection. In this work, we propose a new method to discover collection-adapted ranking functions based on genetic programming (GP). The evolution process of our* Combined Component Approach *(CCA) reuses several components extracted from effective and well-known ranking functions. Our assumption is that these components are representative and meaningful and can be combined for generating a more effective and specialized new ranking function for a given document collection. Experimental results show that our approach was able to outperform in more than 40% classical approaches in two different collections, and also was able to reduce the overtraining, commonly found in machine learning methods, especially genetic programming.*

***Resumo.*** *Com o crescimento da Web, a tarefa de recuperação de informação (RI) transformou-se em um problema extremamente complexo e desafiador. Por isso, diversas funções de ordenação têm sido investigadas ao longo dos anos. No entanto, a maioria delas tem um caráter genérico, isto é, são projetadas para serem efetivas em qualquer coleção. Este trabalho propõe um novo método para descobrir funções de ordenação adaptadas a uma coleção baseado em programação genética (GP). O processo evolutivo da Abordagem de Componentes Combinados (CCA), proposta por este trabalho*[1]*, reutiliza componentes de diferentes funções de ordenação comprovadamente eficazes e conhecidas da literatura de recuperação de informação. Parte-se da hipótese de que estes componentes são individualmente representativos e ricos de significado e podem ser combinado pelo arcabouço GP para a geração de uma nova função de ordenação mais efetiva e especializada para uma determinada coleção. Os resultados experimentais mostram que a abordagem proposta foi capaz de superar em até 40% abordagens clássicas da literatura em duas coleções diferentes e de reduzir o problema do "treinamento exagerado", geralmente encontrado em métodos de aprendizado de máquina, especialmente programação genética.*

[1] Disponível em http://www.dcc.ufmg.br/pos/cursos/defesas/297M.PDF

## 1. Introdução

O desenvolvimento intelectual, social e científico da humanidade está intimamente relacionado ao registro físico do conhecimento humano. Com o advento e o crescimento da *Web* no início da década de 90, considerada como um enorme repositório universal do conhecimento humano, e de outros repositórios de informação, como Bibliotecas Digitais, a tarefa de pesquisar informação tornou-se mais fácil, principalmente devido à disponibilização e ao compartilhamento de um grande volume de informação que passou a estar acessível a milhões de pessoas em todo o mundo a um custo muito baixo. Por outro lado, recuperar informação pertinente a uma necessidade de informação do usuário em meio a bilhões de documentos transformou-se em um problema extremamente complexo e desafiador. Segundo John Battelle [Battelle 2005], a tarefa de busca é um dos problemas mais desafiadores e interessantes de toda a história da Ciência da Computação.

Neste contexto, as máquinas de busca surgiram como ferramentas fundamentais para a tarefa de recuperação de informação (RI) em coleções de documentos. Estas ferramentas são baseadas em modelos de RI, cuja finalidade principal é fornecer o arcabouço teórico para que a máquina de busca possa ser capaz de satisfazer às necessidades de informação dos usuários. Dessa forma, investigar sobre ferramentas de busca e modelos de recuperação de informação que possam atender à necessidade de informação de um usuário de forma mais efetiva, é um tema fundamental nos dias de hoje.

Basicamente os modelos de recuperação de informação tentam encontrar o conjunto de documentos de uma coleção que satisfaça da melhor forma possível à necessidade de informação de um usuário. Isto é, o usuário formula uma consulta, como sendo a expressão de sua necessidade de informação, submete ao processador de consultas de uma máquina de busca e espera como resultado uma lista ordenada de documentos (*ranking*), cuja ordem tenta expressar o grau de relevância de cada documento retornado frente à consulta submetida. A ordem dos documentos é calculada por uma função de similaridade ou função de ordenação (*ranking function*) que atribui um número real a cada documento retornado, representando o grau de similaridade entre o documento e a consulta.

Diferentes funções de ordenação têm sido investigadas ao longo dos anos. No entanto, a maioria delas geralmente tem um caráter genérico, isto é, são projetadas para serem efetivas em qualquer tipo de coleção. Uma função de ordenação mais efetiva é aquela cujas respostas retornadas para uma consulta, aproxima-se do conjunto de documentos relevantes para esta consulta, também chamado por Robertson e Sparck-Jones [Robertson and Sparck-Jones 1976] de conjunto ideal de respostas. O trabalho de Zobel e Moffat [Zobel and Moffat 1998], por exemplo, apresenta mais de 1.000.000 de possibilidades diferentes de calcular uma função de ordenação. Entretanto, ao final dos experimentos, eles concluíram que nenhuma delas é consistentemente efetiva em todas as coleções. Isto é, uma função de ordenação pode ter sucesso em um determinado domínio, mas não ser efetivo em um outro. Outra conclusão importante daquele trabalho é que a exploração exaustiva do espaço de busca das funções de ordenação não é uma solução viável para verificar qual função tem melhor desempenho em uma determinada coleção.

O problema investigado nesta dissertação de mestrado [Almeida 2007] é o da descoberta de funções de ordenação, particularmente usando programação genética (GP - *genetic programming*) [Koza 1992]. Não obstante existir um grande número de funções de ordenação de caráter geral na literatura, este trabalho investiga o problema da geração

de uma função de ordenação mais específica para uma determinada coleção. Uma vez descoberta, esta função poderá ser utilizada na tarefa de recuperação de informação nesta coleção. A escolha de GP deve-se à natureza do problema, isto é, o espaço de possibilidades de funções de ordenação é muito grande e GP tem-se mostrado um método efetivo de otimização em grandes espaços de busca, nas mais diversas aplicações [Koza 1992]. Mais ainda, GP é um método extremamente flexível que permite combinar as evidências disponíveis de forma não-linear e de maneiras não anteriormente previstas, possibilitando gerar funções de ordenação potencialmente mais efetivas do que o estado-da-arte. A outra grande premissa deste trabalho, que o difere de abordagens anteriores, é a proposta de utilização de componentes significantes, advindos de fórmulas efetivas de recuperação de informação, como evidências a serem combinadas pelo arcabouço GP, a partir da hipótese de que esses componentes carregam importante conhecimento que será efetivamente reutilizado pelo arcabouço GP. Como se verá a seguir, os resultados experimentais comprovaram as hipóteses do trabalho de que as estratégias e métodos escolhidos são eficazes na busca de funções de ordenação efetivas.

O restante deste texto está organizado da seguinte forma: a Seção 2 apresenta os objetivos e contribuições deste trabalho; a Seção 3 descreve a abordagem de componentes combinados; a Seção 4 descreve os experimentos realizados e os resultados obtidos; finalmente, a Seção 5 apresenta as conclusões e trabalhos futuros.

## 2. Objetivos e Contribuições

O principal objetivo da dissertação foi contribuir para a descoberta de funções de ordenação mais adaptadas a uma determinada coleção. Para isso, foi proposta uma abordagem usando GP, baseada em componentes bem definidos (e.g., componentes de freqüências dos termos nos documentos e na coleção, componente de normalização) de diferentes funções de ordenação comprovadamente eficazes e conhecidas da literatura de RI. Partiu-se da hipótese de que estes componentes, que em verdade são partes de diversas funções de ordenação, são individualmente representativos e ricos de significado. Aproveitou-se, dessa forma, de todo o conhecimento humano acumulado por trás destas conhecidas funções na geração de uma nova função de ordenação mais especializada para uma coleção. Abordagens anteriores de geração de funções de ordenação usando GP [Fan et al. 2004, Fan et al. 2005, Trotman 2005] utilizam basicamente informações estatísticas brutas da coleção tais como freqüência de termos nos documentos, freqüência de documentos na coleção, tamanho dos documentos e tamanho médio dos documentos. A Abordagem de Componentes Combinados (CCA - *Combined Component Approach*), proposta por este trabalho, é descrita na Seção 3.

As principais contribuições da dissertação, sob os pontos de vista teórico e experimental, foram as seguintes:

- Do ponto de vista teórico, foi proposto um novo arcabouço baseado em GP para a descoberta de funções de ordenação específicas para coleções de documentos. Esse arcabouço é baseado em uma abordagem de componentes combinados que utiliza componentes de funções de ordenação conhecidas e comprovadamente eficazes como a base para todo o processo evolutivo. O uso de componentes ricos e significativos, como terminais do arcabouço GP, ao invés de informações estatísticas mostrou-se bastante efetivo na geração de funções de ordenação especializadas para uma coleção de documentos, com ganhos substanciais sobre as

linhas de base. A abordagem CCA pode ser adaptada para ser aplicada também em outros problemas de RI tais como classificação de texto e recuperação de imagens. Ela também abre novas perspectivas de trabalhos na área mais geral de uso de técnicas de aprendizado de máquina em recuperação de informação. Mais, as análises iniciais dos resultados apresentados na dissertação podem ajudar a elucidar os ganhos obtidos e podem servir de inspiração para outros trabalhos usando GP e RI.

- Do ponto de vista experimental, uma gama extensiva de experimentos atesta a efetividade dessa abordagem em comparação a outras funções de ordenação e outros métodos baseados em GP. De acordo com os experimentos realizados, os resultados obtidos pela abordagem de componentes combinados superaram os resultados de outras abordagens importantes da literatura. Portanto, esses resultados podem também servir de linha de base para novos trabalhos de descoberta de funções de ordenação.

O resultado deste trabalho de pesquisa foi publicado nos anais do ACM SIGIR [Almeida et al. 2007], que é a mais importante conferência internacional da área de recuperação de informação (QUALIS A).

## 3. Abordagem de Componentes Combinados

A Abordagem de Componentes Combinados (CCA - *Combined Component Approach*) proposta por este trabalho é uma abordagem baseada em programação genética para a descoberta de efetivas funções de ordenação. O objetivo principal de CCA é descobrir novas funções de ordenação que sejam mais adaptadas às características de uma coleção de documentos específica. Diferentemente de outros trabalhos de geração de funções de ordenação baseado em GP, a idéia de CCA consiste em examinar importantes funções de ordenação presentes em diferentes modelos e sistemas de RI descritos na literatura, tais como [Buckley et al. 1996, Robertson and Walker 1999, Singhal et al. 1996], e extrair destas funções os componentes de um esquema de ponderação de termos, como aqueles descritos em [Salton and Buckley 1988]. Uma vez identificados, podem ser utilizados como terminais do arcabouço GP, onde serão combinados e utilizados para a geração de novas funções de ordenação mais específicas para uma determinada coleção de documentos.

A idéia da abordagem CCA é descobrir uma função de ordenação $r(q_i, d_j)$ que associe um número real a cada documento $d_j$ da coleção de acordo com a sua similaridade em relação a uma consulta $q_i$. Uma função de ordenação baseada em um sistema de ponderação de termos, que representa a similaridade entre uma consulta $q$ e um documento $d$, pode ser expressa pelo somatório do produto dos pesos de cada termo $t$ de uma consulta em relação a consulta $q$ e ao documento $d$. Esta expressão pode ser simplificada na Equação 1 que define a similaridade entre uma consulta $q$ e um documento $d$ de acordo com a abordagem CCA.

$$sim(q, d) = \sum_{t \in q} w_{tdq} \quad (1)$$

onde $w_{tdq}$ é uma função que retorna o peso de um termo $t$ presente na consulta em relação a um documento $d$ e uma consulta $q$.

Dessa forma, a abordagem CCA tenta descobrir através do arcabouço GP o peso $w_{tdq}$ que faça com que a função de ordenação para uma determinada coleção de documentos se aproxime da solução ótima. Assim como a abordagem proposta por Fan *et al.* [Fan et al. 2004, Fan et al. 2005], CCA usa uma estrutura de dados de árvore para representar o peso $w_{tdq}$. A representação baseada em árvores permite fácil implementação, caminhamento e interpretação. A Figura 1(a) ilustra um indivíduo, $w_{tdq}$, representando um esquema clássico de ponderação conhecido por $tf$-$idf$ [Salton and Buckley 1988], onde $tf$ diz respeito à freqüência de um termo em um documento e $idf$ é uma medida da raridade de um termo na coleção. Os nós-folha assim como em árvores são chamados de terminais e representam unidades de informações básicas que serão utilizadas para criar uma nova fórmula de ponderação para o peso $w_{tdq}$. Terminais são combinados através de funções que são representadas nos nós-internos. Em trabalhos anteriores de descoberta de funções de ordenação baseados em GP [Fan et al. 2004, Fan et al. 2005, Trotman 2005], os terminais sempre refletem informações estatísticas básicas diretamente derivadas de um coleção, tais como freqüência de um termo em um documento ou tamanho de um documento.

A abordagem CCA, no entanto, difere de todas as demais por usar terminais mais significativos. A Figura 1(b) exemplifica um outro indivíduo representando um esquema de ponderação $tf$-$idf$. Neste caso, a informação de $idf$ é por si mesma um terminal e não o resultado de uma combinação de terminais. Nos trabalhos anteriores citados, esta informação é uma sub-árvore que deve ser explicitamente descoberta pelo processo evolutivo do arcabouço GP, podendo até mesmo não ser descoberta. Desse modo, a abordagem CCA tira vantagem do uso de informações previamente conhecidas — componentes de funções de ordenação tais como $idf$ e componentes de normalização pivoteada [Singhal et al. 1996] — para a geração de novas funções de ordenação mais efetivas baseadas nestes componentes, explorando de maneira otimizada e orientada o espaço de busca de soluções.

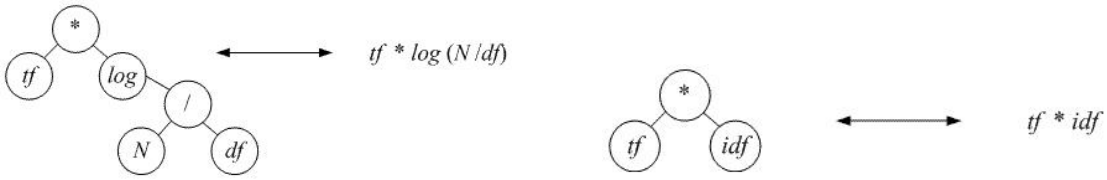



**Figura 1. Um indíviduo $tf$-$idf$ baseado em informações estatísticas (a) e outro baseado na abordagem CCA (b), na qual $tf$ e $idf$ são componentes.**

O uso de terminais mais significativos extraídos de funções de ordenação conhecidas está também fundamentado na definição da propriedade de suficiência [Koza 1992], na qual terminais e funções devem ser capazes de expressar uma possível solução para o problema. Dessa forma, diferentemente dos terminais baseados apenas em informações estatísticas que isoladamente não representam possíveis soluções para o problema, os terminais CCA, por construção, já representam possíveis funções de ordenação ou parte delas e, com isso, têm mais chance de atender a propriedade de suficiência e de convergir em uma solução para o problema investigado.

O arcabouço GP usado em CCA é basicamente um processo iterativo de duas fases: treino e validação. Para cada fase, são selecionados um conjunto de consultas e

documentos da coleção, que são chamados conjunto de treino, para a fase de treino, e conjunto de validação, para a fase de validação. Cria-se uma população inicial de indivíduos que evoluem geração por geração. Em cada fase, é verificada a aptidão de cada indivíduo através da aplicação de uma função de aptidão (*fitness*). Uma vez que cada indivíduo representa um esquema de ponderação, aplicar a função de aptidão significa calcular a função de ordenação para um indivíduo de acordo com o conjunto de documentos e consultas de treino. O valor obtido para a função de aptidão é uma medida de qualidade do *ranking* obtido em relação a relevância dos documentos retornados. Ao final do processo evolutivo, escolhe-se o melhor indivíduo considerando o desempenho nas fases de treino e validação.

Os terminais, as funções, as funções de aptidão (MAP e FFP4) e os métodos de seleção dos melhores indivíduos ($\text{SUM}_\sigma$ e $\text{AVG}_\sigma$) utilizados pela abordagem CCA, assim como os melhores indivíduos descobertos, estão descritos em [Almeida 2007].

## 4. Experimentos

Para avaliação da abordagem de componentes combinados proposta neste trabalho, foram utilizadas as coleções TREC-8 [Voorhees and Harman 1999] e WBR99[2].

### 4.1. Avaliação dos Experimentos

Os resultados obtidos nos experimentos realizados com a abordagem CCA foram comparados com outras três abordagens: (i) Okapi BM25 [Robertson and Walker 1999] (usando os parâmetros $k_1$=1.2, $k_3$=1000, $b$=0.75), (ii) modelo de espaço vetorial clássico com esquema de ponderação $tf\text{-}idf$, e (iii) a abordagem GP proposta por Fan *et al.* em [Fan et al. 2004] (denominada de FAN-GP neste trabalho). Foram escolhidas, como linha de base de comparação, as duas melhores abordagens para cada coleção. Para a TREC-8, BM25 e FAN-GP. Já para a coleção WBR99, TF-IDF e FAN-GP. Para a coleção TREC-8, foram utilizadas 20 consultas para treino, 10 para validação e 20 para testes. Assim como, para a coleção WBR99, foram utilizadas 20 consultas para treino, 10 para validação e 19[3] para testes. O conjunto de testes foi utilizado para testar a efetividade de CCA contra os demais métodos. Ao final do processo evolutivo, a seleção das melhores funções de ordenação descobertas foi realizada levando em consideração o desempenho obtido nas fases de treino e validação conjuntamente. Em ambas as coleções, a melhor função de ordenação foi escolhida através do método $\text{SUM}_\sigma$.

### 4.2. Resultados

A Tabela 1 apresenta detalhadamente os resultados obtidos pela abordagem CCA para as coleções TREC-8 e WBR99. A significância estatística dos ganhos apresentados contra os *baselines* foi comprovada através do teste $t$. Como pode ser visto, CCA alcança melhores resultados do que BM25 e FAN-GP. Na precisão média, CCA atingiu 16,40%, superando BM25 em 40,87% e FAN-GP em 14,00%, para a TREC-8, usando a função de aptidão MAP. Para a coleção WBR99, usando a função de aptidão FFP4, também foram obtidos ganhos em relação às abordagens utilizadas como linha de base para comparação. CCA obteve o valor de 16,68% na precisão média. Este valor supera em 21,67% TF-IDF e em 24,85% FAN-GP. Pode ser observado também que FAN-GP não conseguiu superar TF-IDF para esta coleção.

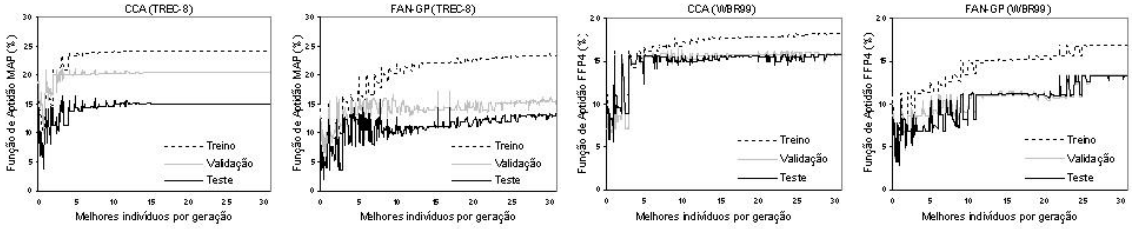
[2]Disponível em http://www.linguateca.pt/Repositorio/WBR-99/

[3]Não existe disponível a informação de relevância para a consulta 35.

| Nível | TREC-8 | | | | | WBR99 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Baselines | | CCA | | | Baselines | | CCA | | |
| | BM25 | FAN-GP | *Precisão* | *Ganho sobre BM25* | *Ganho sobre FAN-GP* | TF-IDF | FAN-GP | *Precisão* | *Ganho sobre TF-IDF* | *Ganho sobre FAN-GP* |
| Em 5 doctos | 27,000 | 27,000 | 32,000 | +18,52% | +18,52% | 23,157 | 35,790 | 36,842 | +59,10% | +2,94% |
| Em 10 doctos | 29,000 | 25,500 | 31,500 | +8,62% | +23,53% | 26,315 | 32,632 | 32,632 | +24,00% | 0,00% |
| Em 15 doctos | 25,333 | 24,667 | 27,000 | +6,58% | +9,46% | 30,175 | 29,474 | 29,123 | -3,49% | -1,19% |
| Em 20 doctos | 23,750 | 22,750 | 25,000 | +5,26% | +9,89% | 32,105 | 26,842 | 27,368 | -14,75% | +1,96% |
| Em 30 doctos | 20,167 | 21,333 | 22,167 | +9,92% | +3,91% | 25,263 | 21,579 | 25,965 | +2,78% | +20,33% |
| Em 100 doctos | 15,050 | 15,150 | 16,050 | +6,64% | +5,94% | 13,421 | 10,737 | 13,632 | +1,57% | +26,96% |
| Em 200 doctos | 11,900 | 11,775 | 12,025 | +1,05% | +2,12% | 9,000 | 7,553 | 8,790 | -2,34% | +16,38% |
| Em 500 doctos | 7,410 | 7,160 | 7,650 | +3,24% | +6,84% | 3,726 | 3,516 | 3,737 | +0,28% | +6,29% |
| Em 1000 doctos | 4,910 | 4,505 | 5,000 | +1,83% | +10,99% | 1,.968 | 1,879 | 2,005 | +1,87% | +6,73% |
| **R-precision** | 16,116 | 17,703 | 20,449 | +26,89% | +15,51% | 20,607 | 19,830 | 22,294 | +8,19% | +12,43% |
| **Precisão Média** | 11,643 | 14,388 | 16,402 | +40,87% | +14,00% | 13,710 | 13,361 | 16,681 | +21,67% | +24,85% |
| *Confiança* | | | | *94,05%* | *98,19%* | | | | *96,96%* | *96,04%* |

**Tabela 1. Precisão obtida pela abordagem CCA para TREC-8 e WBR99**

A Figura 2 mostra o processo evolutivo CCA ao longo de 30 gerações, considerando os 20 melhores indivíduos, para as coleções TREC-8 e WBR99. Para cada geração, foram avaliados, de acordo com a função de aptidão utilizada, os 20 melhores indivíduos. Como pode ser observado em (a) e (c), as funções descobertas pela abordagem CCA convergem mais rapidamente do que as descobertas pela abordagem FAN-GP, (b) e (d). As figuras também mostram que as curvas CCA comportam-se de modo mais similar. Isto é, apesar do fato dos conjuntos de treino, validação e teste apresentarem diferentes valores de aptidão, as curvas de validação e teste tendem a seguir o comportamento da curva de treino. Na abordagem FAN-GP, as curvas de validação e teste não seguem o comportamento da curva de treino, indicando *overfitting*. Isto pode ser observado tanto na coleção TREC-8 quanto na coleção WBR99. Outro indicador da ocorrência de *overfitting* é a distância entre as curvas de treino e as demais curvas. Ou seja, quanto maior a distância entre as curvas, maior o *overfitting*. Observando as figuras, pode ser comprovado que as curvas CCA em (a) e (c) estão mais próximas uma das outras do que as curvas da abordagem FAN-GP em (b) e (d), indicando que na abordagem CCA o problema do *overfitting* ocorre de modo menos acentuado do que em FAN-GP.




(a) (b) (c) (d)

**Figura 2. Processo evolutivo CCA considerando os 20 melhores indivíduos em 30 gerações para as coleções TREC-8 e WBR99.**

Durante a pesquisa, foi realizada também uma etapa de análise das funções geradas que basicamente procurou investigar as causas para o bom desempenho dessas funções. A melhor função de ordenação descoberta em cada um dos experimentos executados foi examinada para se verificar a ocorrência de cada terminal, visando descobrir os mais freqüentes. Notou-se então que os terminais mais freqüentes em cada função descoberta eram parte de funções reconhecidamente efetivas para cada coleção, indicando que a abordagem CCA foi capaz de escolher os terminais mais apropriados para as características de cada coleção e utilizá-los de forma efetiva para gerar funções de ordenação melhores do que aquelas conhecidas.

## 5. Conclusões e Trabalhos Futuros

Neste trabalho, foi apresentada a abordagem CCA para a geração de funções de ordenação. Os resultados mostraram que CCA melhorou a efetividade da tarefa de recuperação de informação quando comparada às abordagens tradicionais e outra abordagem GP (FAN-GP) [Fan et al. 2004], utilizando as coleções TREC-8 e WBR99. Examinando o processo evolutivo da abordagem CCA, foi observado que as funções descobertas convergiram mais rapidamente do que FAN-GP. Além disso, CCA também conseguiu reduzir o problema do "treinamento exagerado", no qual as funções ficariam muito especializadas para o conjunto utilizado no treino. Os resultados obtidos permitem a conclusão de que o uso de terminais significativos, como os componentes extraídos de outras funções de ordenação, ao invés de usar simplesmente informações estatísticas básicas, melhora a qualidade do processo de descoberta de funções de ordenação usando GP.

Como trabalhos futuros, pretende-se evoluir a abordagem para utilizar evidências de documentos Web (ligações, estrutura etc.), descobrir funções de ordenação específicas para grupo de consultas (navegacionais, informacionais etc.) e realizar uma análise mais detalhada das funções de ordenação descobertas.

## Referências

Almeida, H. M. (2007). Uma abordagem de componentes combinados para a geração de funções de ordenação usando programação genética. Dissertação de Mestrado, DCC, UFMG. Disponível em http://www.dcc.ufmg.br/pos/cursos/defesas/297M.PDF.

Almeida, H. M., Gonçalves, M. A., Cristo, M., and Calado, P. (2007). A combined component approach for finding collection-adapted ranking functions based on genetic programming. In *Proc. of the 30th ACM SIGIR*, pages 399–406, Amsterdam, Netherlands.

Battelle, J. (2005). *A Busca*. Elsevier, Rio de Janeiro.

Buckley, C., Singhal, A., and Mitra, M. (1996). New retrieval approaches using smart: TREC 4. In *Proc. of TREC-4*, pages 25–48, Gaithersburg, MD. NIST Special Publication 500-236.

Fan, W., Gordon, M. D., and Pathak, P. (2004). A generic ranking function discovery framework by genetic programming for information retrieval. *Information Processing and Management*, 40(4):587–602.

Fan, W., Gordon, M. D., and Pathak, P. (2005). Genetic programming-based discovery of ranking functions for effective web search. *Journal of Management Information Systems*, 21(4):37–56.

Koza, J. R. (1992). *Genetic Programming: On the programming of computers by natural selection*. MIT Press, Cambridge.

Robertson, S. E. and Sparck-Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146.

Robertson, S. E. and Walker, S. (1999). Okapi/keenbow at TREC-8. In *Proc. of TREC-8*, pages 151–162, Gaithersburg, MD. NIST Special Publication 500-246.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

Singhal, A., Buckley, C., and Mitra, M. (1996). Pivoted document length normalization. In *Proc. of the 19th ACM SIGIR*, pages 21–29, Zurich, Switzerland.

Trotman, A. (2005). Learning to rank. *Information Retrieval*, 8(3):359–381.

Voorhees, E. M. and Harman, D. (1999). Overview of the eighth Text REtrieval Conference (TREC-8). In *Proc. of TREC-8*, pages 1–24, Gaithersburg, MD. NIST Special Publication 500-246.

Zobel, J. and Moffat, A. (1998). Exploring the similarity space. *SIGIR Forum*, 32(1):453–490.